Predict the survival of the passengers aboard RMS Titanic

ABSTRACT

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. The RMS Titanic was the largest ship afloat at the time it entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. The Titanic was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, her architect, died in the disaster. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. In this project, I analysed of what sorts of people were likely to survive. **In particular, we tried to apply the tools of machine learning that is "Logistic Regression" to predict which passengers survived the tragedy.**

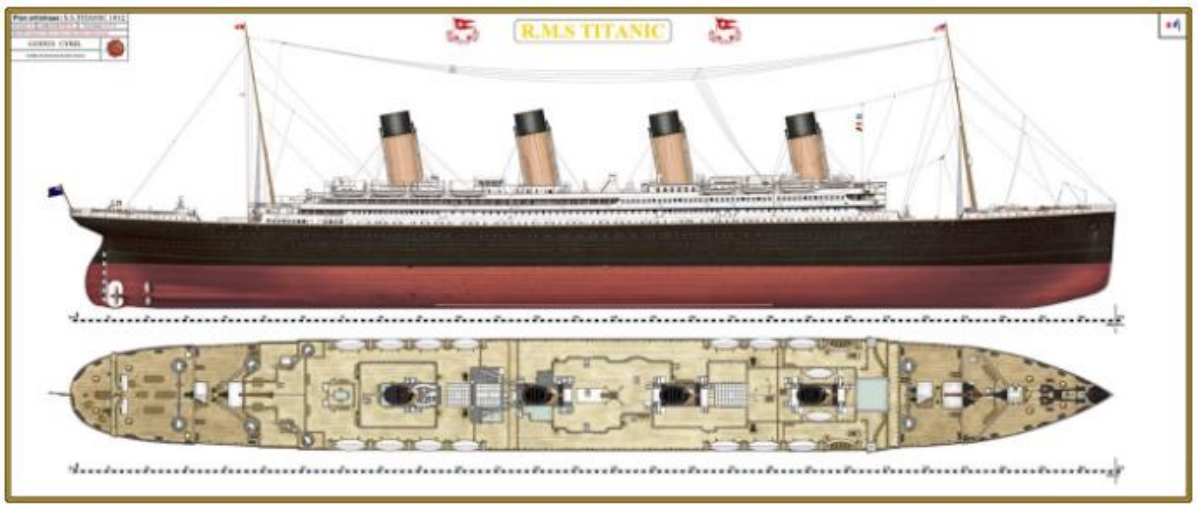

Fig 1: RMS Titanic

INTRODUCTION

Only one disaster the last 100 years remains etched in people's awareness and imagination a century after the accident: TITANIC. 35 movies have been made with themes from the accident, several books have been written and there are around 15 million listings on Google. One hundred years ago the highly travelled route across the North Atlantic was very lucrative. Ship builders constructed ocean liners that grew larger and faster with each

generation. The driving forces in the market were speed, size and luxury. Among the shipping lines that were in an endless war for dominance was the White Star Line.

In 1902 the White Star Line was sold to an American financier J. Pierpont Morgan, who became interested in shipping companies because of the growing passenger traffic to North America. It was Morgan's money that allowed the dream of the "Olympic Class" liners Olympic, Titanic and Britannic to come true. When the Titanic left Queenstown Ireland in the afternoon of Thursday 11th April 1912 she was following what had been accepted since 1899 as the outward-bound route for mail steamers to the United States. The selection of the route was based on the importance of avoiding areas where fog and ice were prevalent at certain seasons, without lengthening the passage across the Atlantic. On 15th April, at 2.20 am, Titanic broke apart and foundered.

Machine learning algorithms are applied to make a prediction which passengers survived at the time of sinking of the Titanic. Features like ticket fare, age, sex, class will be used to make the predictions. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data. Using the machine learning algorithms, survival is predicted on different combinations of features.

The objective is to perform data analytics to mine various information which are available in the dataset and to know effect of each field on survival of passengers by applying analytics.

DATA ANALYTICS AND ITS CATEGORIES

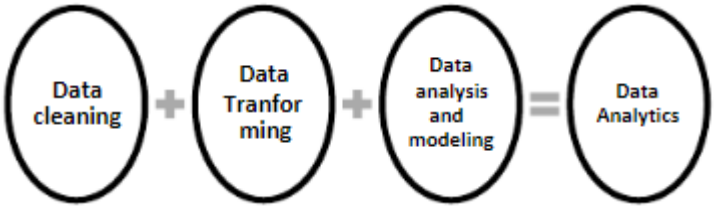

Fig 2: Data Analytics diagram

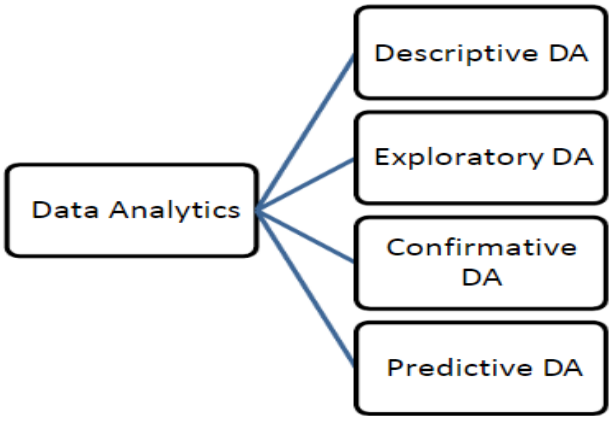

Fig 3: Categories of Data Analytics

PROCESS FLOW

There is a step by step approach to choose a particular model for the current problem. We need to decide whether a particular machine learning model is suitable for our problem or not. Here we can see process flow being followed.

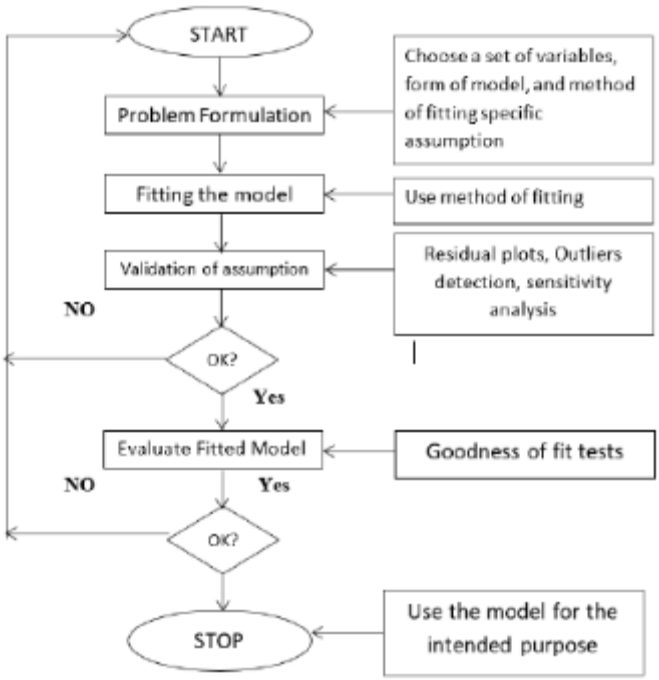

Fig 4: Process of fitting a Machine Learning Model

DESCRIPTION OF DATA

In Python info() function is used to find structure of dataset that we have in csv file. Below there is a snippet of output of we got by executing info() in Python.

```
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Fig 5: Structure of input Dataset

Attribute	Description	Factors
PassengerId	ID of a passenger	
Survived	Survival of passenger	0 = NO, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Name of the passenger	
Sex	Sex	Male/Female
Age	Age of the passengers in years	
SibSp	# of siblings and spouses abord the Titanic	
Parch	# of parents/children abord the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port from where passenger embarked. C for Cherbourg, Q for Queenstown, S for Southampton	C, Q, S

Fig 6: Description of attributes in our Dataset

EXPLORATORY DATA ANALYSIS

We have performed "Exploratory Data Analysis" for our problem in the first stage. In exploratory data analysis dataset is explored to figure out the features which would influence the survival rate. The data is deeply analysed by finding a relationship between different attribute and survival.

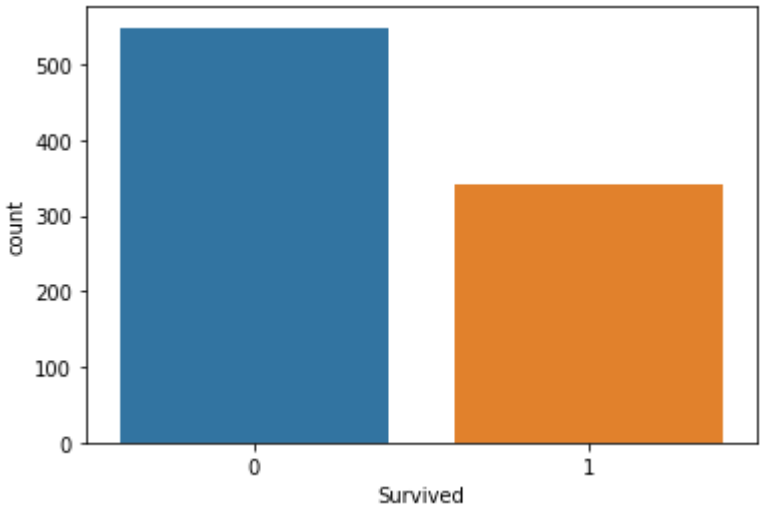

Fig 7: Number of people survived or not survived

Age verses Survival – Figure, 8 shows how survival rate is affected by age. If the value of age is less then chances of survival are more and vice-versa.

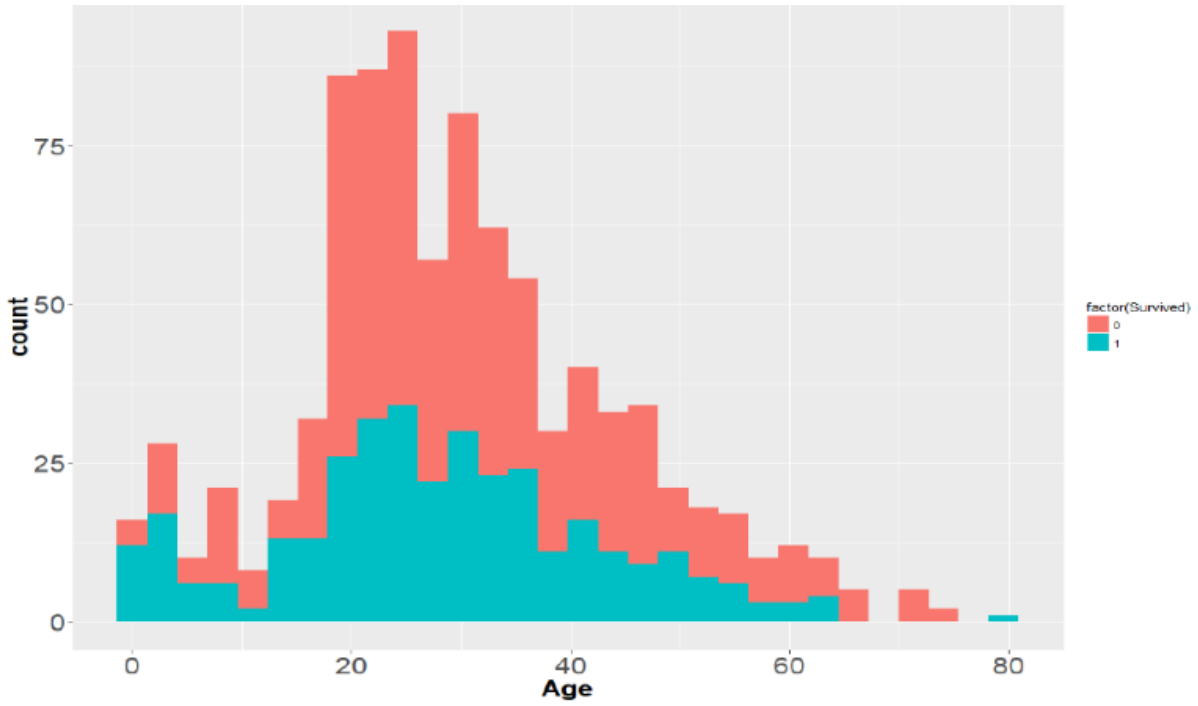

Fig 8: Age V/s Survival

Sex verses Survival - From Fig. 9 it is clear that females are more likely to survive than males. **We calculated that survival rate of female and male are 74.20382% and 18.89081% respectively.**

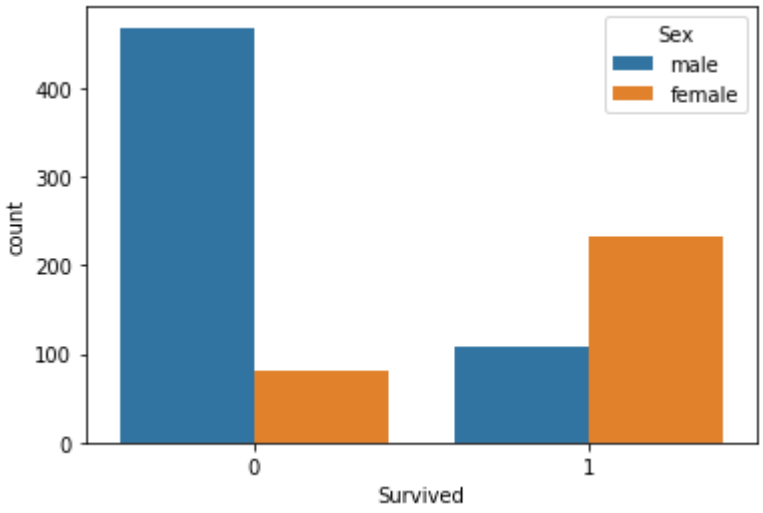

Fig 9: Sex V/s Survival

Pclass verses Survival – From figure, 10 it is clear that passengers travelling in the lower or 3rd class did not tend to survive much. On the other hand those were travelling in the 1st class tend to survive more than 2nd and 3rd class passengers.

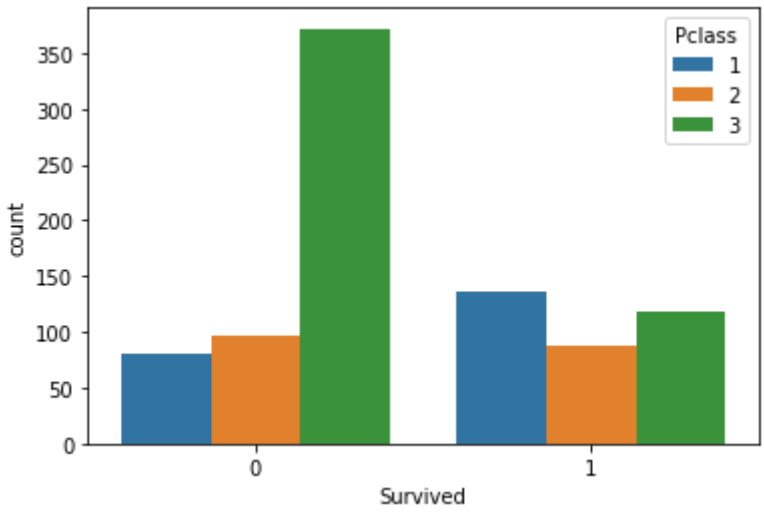

Fig 10: Pclass V/s Survival

In similar way relationship between other attributes like fare, cabin, title, family, Pclass, Embarked and survival is found. In this way we will be able to decide emphasis of each attribute on survival of passenger.

DATA WRANGLING AND CLEANING

Before applying any type of data analytics on the dataset, the data is first cleaned. There are some missing values in the dataset which needs to be handled. Figure, 11 displays the number of missing or null values for a particular attribute in the dataset. We can see that the attribute "Cabin" has the highest number of missing values. "Yellow" band shows the number of missing values. So, before creating our model we need to either fill those fields with some dummy values or the mean value of that particular attribute. Here, in my model we just dropped that row which contains any null or missing values.

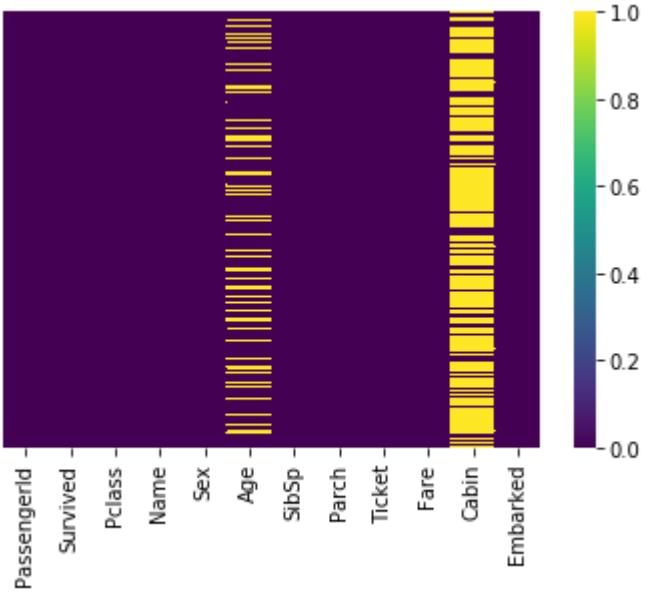

Fig 11: Number of null values before data cleaning

After performing data wrangling we got the cleaned dataset which can be visualize from figure 12, which shows there is no null or missing values in the dataset.

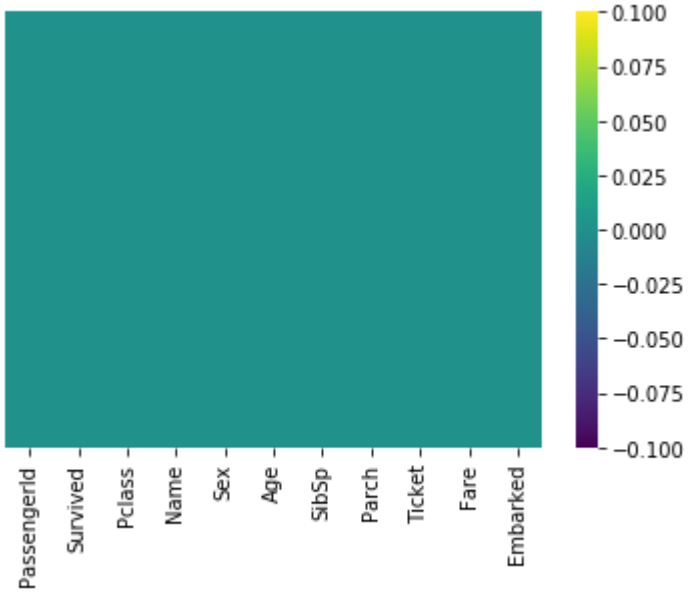

Fig 12: Number of null values after data cleaning

Also, there were many string and character values in the dataset which needs to convert into its corresponding categorical values in order to apply machine learning algorithm that is **"Logistic Regression"**. "Sex" attribute has been changed to "male", which contains 1 if the passenger is male and 0 for female. "Embarked " attribute has been changed to "Q" & "S", which also contains 0s and 1s. Similarly, "Pclass" has been changed to "2" & "3", which also contains 0s and 1s.

As, for predicting the survival of passenger we actually don't need passenger's PassengerID, Name, Ticket number so, we can drop those attributes or that particular column. After completion of data wrangling we got our final dataset which is shown in the figure 13.

	Survived	Age	SibSp	Parch	Fare	male	Q	S	2	3
0	0	22.0	1	0	7.2500	1	0	1	0	1
1	1	38.0	1	0	71.2833	0	0	0	0	0
2	1	26.0	0	0	7.9250	0	0	1	0	1
3	1	35.0	1	0	53.1000	0	0	1	0	0
4	0	35.0	0	0	8.0500	1	0	1	0	1

Fig 13: Final Dataset showing top 5 rows

LOGISTIC REGRESSION

Logistic regression is the technique which works best when dependent variable is dichotomous (binary or categorical). The data description and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables is done with the help of logistic regression. It is used to solve binary classification problem, some of the real life examples are spam detection-predicting if an email is spam or not, health-Predicting if a given mass of tissue is benign or malignant, marketing-predicting if a given user will buy an insurance product or not.

PREDICTION

For training our model, we divided the whole dataset into 60 – 40 ratio and defined X and Y axis variable. We have assigned "Survived " attribute which is a dependent variable to Y-axis and all other attributes accept survived which are independent attributes has been assigned to X-axis. Then, we performed the prediction on our dataset using the model called "Logistic Regression".

MODEL EVALUATION

The accuracy of the model is evaluated using "confusion matrix". A confusion matrix is a table layout that allows to visualize the correctness and the performance of an algorithm.

Confusion Matrix

.

A confusion matrix is a method to verify how accurately the classification model works. It gives the actual number of predictions which were correct or incorrect when compared to the actual result of the data. The matrix is of the order N*N, here N is the number of values. Performance of such models is commonly evaluated using the data in the matrix.

Figure 14 shows the confusion matrix which we got for the model.

```
array([[139,  31],
       [ 31,  84]], dtype=int64)
```

Fig 14: Confusion Matrix

Accuracy

It gives the measure of percentage of correct prediction done by the model/algorithm. The best value is "1.0" and the worst value is "0.0".

In our model we got an accuracy score of 78.24%, which is not so bad.

Along with the confusion matrix and accuracy sore we have also generated classification report of our model which is shown in figure 15 below. It contains precision values and f1 - score.

```
'              precision    recall  f1-score   support\n\n           0       0.82      0.82      0.82       170\n           1
0.73      0.73      0.73       115\n\n    accuracy                           0.78       285\n   macro avg       0.77      0.77
0.77       285\nweighted avg       0.78      0.78      0.78       285\n'
```

Fig 15: Classification report

Conclusion

Data cleaning is the first step while performing data analysis. Exploratory data analytics helps one to understand the dataset and the dependency among the attributes. EDA is used to figure out the relationship between the features of the dataset. This is done by using various graphical techniques. The one which we have used are countplot, heatmap etc.

There is high influence of age on survival. We can see from the above figure 8 that as age increases survival decreases. It can be seen that survival rate of female is very high (approx. 74%) and survival rate of male is very low. This fact can also be verified by extracting titles (Mr, Mrs, Ms etc) from name column. Survival rate with title Mr. is approximately 16% while survival rate for Mrs. is 79%.

We found that Passengers who were travelling in first class is more likely to survive.

It is clearly stated that the accuracy of the models may vary when the choice of feature modelling is different. Ideally logistic regression and support vector machine are the models which give a good level of accuracy when it comes to classification problem.